

On the Functional Load of Syntactic Phenomena in Three-dimensional Linguistics: A Quantitative Survey in Modern Korean

Sang-Oak Lee

To extend the meaning of "functional load" to the volume (or quantity) of function, or the frequency of usage, I named this study "three-dimensional linguistics": because each rule is not flatly in the same volume but differs in the quantity of usage (or occurrence). One may not know the degree of importance and/or plausibility of rules *a priori* before they are measured in the way as I suggest here. I would like to apply this new definition of "functional load" to the field of syntax as well as phonology. After describing the outline of Korean syntactic phenomena in section 2, I would like to present the results of counting their occurrences in a corpus of sentences. Observation, interpretation, and some analysis of data will follow. In measuring syntactic phenomena, we may notice what sorts of rules most frequently occur in a given language. The importance of such an investigation becomes clear when considering pedagogical purposes and machine translation, among other possible fields. For instance, a teacher can set the order of teaching rules of grammar after considering their frequency. And an engineer of machine translation programs can design the most economical processor by applying this result to his/her implementation.

1. Introduction

This is not a theoretical article, but a report of some interesting facts that linguists have long neglected, but that can be utilized in many areas of linguistics and applied fields.

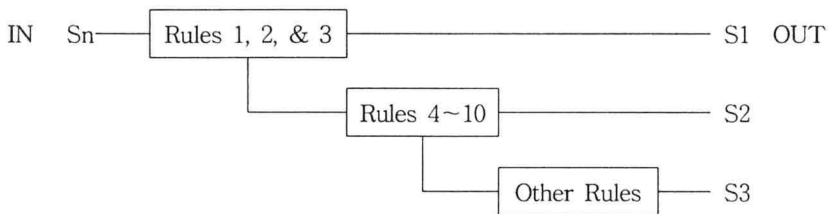
"Functional load" is a term for measuring the frequency of phonological

oppositions.¹ In this paper, however, I would like to extend the sense of this term, in other words.

To date, linguists have used this term mainly for counting the occurrence of a certain phoneme. Yet we may also measure the occurrence of rules or, if the term “rules” is not appropriate, then the expression “phenomena” can be used. As for phenomena, not only phonetic and phonological ones as in S. O. Lee (1990), but also syntactic ones can be measured. This paper aims at what I call merger and analysis (M & A, not in the sense of merger and acquisition) between phonology and syntax.

There might be arbitrariness in selecting and collapsing of phenomena (or rules). However, this problem has overcome as much as possible considering all the grammatical information in synchronic (and even diachronic, if necessary) changes in Korean.

Thus, the most frequent syntactic rules are checked in the first cycle of translation and the subsequent rules may be applied in turn.



〈Figure 1〉

Figure 1 shows a schematic process of economized machine translation. When a certain sentence is input to the first filter consisting of rules 1~3, if the target sentence does not contain any other rules than 1~3, then this sentence is completely processed and outputted as S1. If not, the given sentence should go down to the second filter in order and undergo a similar procedure.

¹ The term ‘the functional load’ as defined by the Prague School is useful for measuring the frequency of phonological oppositions. King (1965, 67) discusses its role in historical linguistics. Many scholars including Meyerstein (1970) have restricted the use of this term within phonology, however.

2. Listing Rules

In order to get information about quantitative aspects of Korean sentence structures, I collected most cases of Korean syntactic rules that have been presented in many Korean syntactic papers written in the framework of earlier transformational grammar. I took this 'earlier' stage because it is still conservative in the field of teaching pedagogical grammar. I also selected some sample Korean sentences considering the balance of different genres.² I investigated them based on the list of these rules as shown below.

A List of Korean Syntactic Rules

(1) Pronominalization

A pronoun is used instead of repeating nouns.

- a. Personal Pronoun : na 'I', uli 'we', ce(huy) '(humble) we', ne 'you', nehuy 'you'(pl.)
- b. Indefinite Pronoun : nwugwu 'who', amwu(gay, get) 'anyone, anything', motwu 'all', mwuet 'what'.
- c. Reflexive Pronoun : ce, caki 'self', tangsin '(honorific) self'

(2) Equi-NP Deletion

In a complex sentence where a main clause and an embedded clause contain the same noun phrase, the noun phrase in the embedded clause is deleted.³

² Professor Martin Kay mentioned to me in the question and answer session following his '91 UCSC/LSA Institute lecture on machine translation that the METAL system for German-English translation by Siemens introduced an idea similar to mine. This system is essentially based on the transfer approach, it is written in the LISP programming language, and it represents one of the most advanced operational MT systems at the present time.

³ The terminologies used in our definition of syntactic phenomena are rather old-fashioned since we take the traditional or earlier version here to keep the neutral position out of currently debated issues in syntax. Otherwise, depending on the theories of different claims definitions may vary and this kind of statistical research cannot be conducted. It is not a matter of concern how to define the phenomena and which 'fashion' is taken, but we have to identify the existence of syntactic phenomena in a certain way and factually count their occurrence without fail. The present version of syntactic rules simply represent such a certain way.

e.g. [chelswu-ka [chelswu-ka cip-ey ka-ki]-lul wenha-n-ta.]
 Chelswu-Nom C.-Nom home-Loc go-Nmz-Acc want-Pres-Dec
 'Chelswu wants to go home.'
 → [chelswu_i-ka [O_i cip-ey ka-ki]-lul wenha-n-ta.]

(3) Topicalization

When nouns representing old information appear in sentence-initial or subject position, they have the function of topics. The topic marker is *-un/nun*.

e.g. sonamwu-nun sangnokswu-(i)-ta
 pinetree-Top evergreen tree-(be)-Dec
 'A pinetree is an evergreen tree.'

(4) Relativization

An underlying sentence with adnominal suffixes like *-un*, *-ul* qualifies the following noun phrase, which is underlyingly a constituent of relative clause.

e.g. [ecec ei ilk-un] chayk-un cham caymiit-te-la.
 yesterday read-Rel book-Top very interesting-Retro-Dec
 'The book that I read yesterday was very interesting.'

(5) NP - Complementation

An underlying sentence with adnominal suffixes like *-un*, *-ul* qualifies the following noun phrase, whose contents are represented by the embedded sentence.

e.g. [pemin-i tasi hyencang-e nathana-n] sasil-i swusacin-ul
 criminal-Nom again scene-Loc appear-Comp fact-Nom investigator-Acc
 kincang-sikhi-ess-ta.
 tense-Caus-Past-Dec
 'Investigators were tensed by the fact that the criminals reappeared at the scene (of the crime).'

(6) VP - Complementation

An underlying sentence with connecting suffixes like *-a/e*, *-key*, *-ci*, *-ko* complements the following verb phrase composed of an auxiliary verb.

e.g. yenghuy-nun [takchi-nun taylo ilk-e] tay-nun supkwan-i
 Yenghuy-Nom come close-Comp as read-Comp repeat-Comp habit-Nom

iss-ta.

have-Dec

'Yenghuy has a habit of reading voraciously any book at hand.'

(7) Nominalization

An underlying sentence with nominalizing suffixes like *-um*, *-ki* has the syntactic function of a noun phrase in a complex sentence structure.

e.g. salam-tul-i ttena-ki sicakhay-ss-ta.

people-Pl-Nom depart-Nmz start-Past-Dec

'People began to depart.'

(8) Conjunction

a. Coordinate Conjunction

Coordinating suffixes like *-ko*, *-umye*, *-una*, *-kena*, *-ciman* are used to combine two underlying sentences into one sentence.

Here, two clauses are equivalent and independent in meaning.

e.g. kkoch-i phi-ko, say-ka wu-n-ta.

flower-Nom bloom-and bird-Nom sing-Pres-Dec

'Flowers bloom and birds sing.'

b. Subordinate Conjunction

Subordinating suffixes like *-ko*, *-umye*, *-umyense*, *-ko(se)*, *-ca*, *-camaca*, *-e(se)*, *-kilo(seni)*, *-eto*, *-ilato*, *-telato*, *-tunci*, *-una*, *-kena*, *-untul*, *-ulcilato*, *-atcca*, *-ulmangceng*, *-taman*, *-ciman*, *-myen*, *-esenun*, *-ketun*, *-attentul*, *-toy*, *-nuntey*, *-kenman*, *-uncuk*, *-unpa*, *-ulccintay*, *-keniwa*, *-telani*, *-ule*, *-ulye(ko)*, *-koca*, *-key*, *-tolok*, *-taka*, *-tut*, *-ulssulok*, *-eya* are used to combine two underlying sentences into one sentence. Here, the main clause and the subordinate clause are related in meaning.

e.g. chelswu-ka nuc-ke o-ase yenghuy-nun hwa-ka na-ss-ta.

C.-Nom be late-Comp come-Conj Y.-Nom anger-Nom get-Past-Dec

'Because Chelswu was late, Yenghuy got angry.'

(9) Conjunction Reduction

When there are identical constituents in both clauses of a coordinate sentence, either constituent of them is deleted.

e.g. hyeng-i nolay-lul pwull-ess-ko, tongsayng-i

elder brother-Nom song-Acc sing-Past-Comp younger brother-Nom
nolay-lul pwull-ess-ta.

song-Acc sing-Past-Dec

'Elder brother sang a song and younger brother sang a song.'

→ hyeng-kwa tongsayng-i nolay-lul pull-ess-ta.

-Com

'Elder and younger brothers sang a song.'

(10) Negation

Adverbs and predicates with the meaning of negation in an underlying sentence are used to negate the sentence.

- a. *an* Negation : negative adverb *ani(an)* or negative predicate *ani(ha)ta* is used.
- b. *mot* Negation : negative adverb *mot* or negative predicate *mothata* is used.
- c. *malta* Negation : negative predicate *malta* is used to negate an imperative sentence or a propositive sentence.

(11) Causativization

a. Short Causativization

Causative suffixes like *-i / hi / li / ki / (i)u / ku / chu-* are attached to the stem of the adjectives, intransitive verbs, and transitive verbs with the predicate function in an underlying sentence, resulting in causative verbs. Here, a new causer is introduced and the subject as a causee in the sentence becomes the direct or indirect object.

e.g. *tongsayng-i sakwa-lul po-ass-ta.*

younger brother-Nom apple-Acc see-Past-Dec

'Younger brother saw an apple.'

→ *hyeng-i tongsayng-ekey sakwa-lul po-i-et-ta.*

elder b.-Nom younger b.-Dat apple-Acc see-Caus-Past-Dec

'Elder brother showed younger brother the apple.'

b. Long Causativization

An adverbializing suffix *-key* and an auxiliary predicate *HA-* comes after the stem of adjectives, intransitive verbs, and transitive verbs.

e.g. *tongsayng-i sakwa-lul po-ass-ta.*

younger brother-Nom apple-Acc see-Past-Dec

'Younger brother saw the apple.'

→ hyeng-i tongsayng-ekey sakwa-lul po-key hay-ss-ta.
 elder b.-Nom younger b.-Dat apple-Acc see-Comp let-Past-Dec
 'Elder brother made/let younger brother see the apple.'

(12) Passivization

a. Short Passivization

Passive suffixes *-i/hi/li/ki-* are attached to the stem of transitive verbs.

e.g. swunkyeng-i pemin-ul cap-ass-ta.
 policeman-Nom criminal-Acc catch-Past-Dec
 'A policeman caught a criminal.'

→ pemin-i swunkyeng-eke cap-hi-et-ta.
 criminal-Nom policeman-Dat catch-Pass-Past-Dec
 'A criminal was caught by a policeman.'

b. Long Passivization

A connecting suffix *-e/a* and an auxiliary predicate *-ci-* comes after the stem of verbs.

e.g. mwulkoki-ka cal cap-a ci-n-ta.
 fish-Nom well catch-Comp become-Pres-Dec
 '(lit.) Fish are caught easily.'

c. Special Passivization :

A suffix *-toy-* is attached to the noun.

e.g. cengpwu-ka i-kos-ul kaypalhay-ss-ta.
 government-Nom this place-Acc develop-Past-Dec
 'The government has developed this place.'
 → i-kos-i cengpwu-e uyhay kaypal-toy-ess-ta.
 this place-Nom government-Agent develop-Pass-Past-Dec
 'This place was developed by the government.'

(13) Honorification

The prefinal suffix *-si-* is attached to the stem of predicates to denote the honorification of the subject.

e.g. sensayngnim-i ka-si-n-ta
 teacher-Nom go-Hon-Pres-Dec
 'My (honorable) teacher goes.'

(13) and (14) are not necessarily syntactic phenomena. They could be semantic/pragmatic ones. However, I inclusively enumerate them here since they are somehow relevant to the syntactic phenomena of the sentence as well.

(14) Pluralization

The plural suffix *-tul* is attached usually to nouns, representing the plurality of nouns.

e.g. *sensayngnim-tul*
teacher-Pl

(15) Question Formation

Interrogative suffixes like *-(nu)nya*, *-(nu)nka*, *-o*, *-(upni)kka*, *-e*, *-yo* appear in sentence-final position representing the question form.

e.g. *ney-ka ka-nunya?*
you-Nom go-Inter
'Do you go?'

3. Results of Counting Occurrences

The fifteen items⁴ introduced in the previous section were checked and marked in approximately 300 passages chosen from a Korean composition textbook and university code, etc. The markings were accumulated and calculated in percentage.

The Functional Load of Syntactic Rules :

A Quantitative Survey in Modern Korea

(%)

1. NP-Complementation	NP _i -은,을 NP _j	18.93	
2. Subordinate Conjunction		17.25	
3. Relativization	(NP _i)-은,을 NP _i	15.58	
			-51.76
4. Topicalization		10.14	

⁴Of course, the fifteen items are non-exhaustive and differently classifiable depending on a chosen fashion. Also, these fifteen items are consisting of some twenty subclasses in our version.

5. Coordinate Conjunction	8.28	
6. VP-Complementation -아, 게, 지, 고	7.37	
7. Pronominalization (Personal-)	5.80	
8. Nominalization	3.58	
9. <i>toy</i> -Passivization	2.70	
10. <i>an</i> Negation	1.97	
		-91.60
11. Pluralization	1.93	
12. Interrogation	1.71	
13. Long Passivization - 어지 -	1.42	
14. Indefinite Pronominalization	1.24	
15. Short Passivization	0.51	
		-98.41
16. <i>mot</i> Negation	0.51	
17. Reflexivization	0.44	
18. Short Causativization	0.36	
19. <i>mal</i> -Negation	0.11	
20. Long Causativization - 게 하 -	0.07	
21. Conjunction Reduction	0.07	
22. Honorification	0.07	

4. Observation and Interpretation

As shown above, the most frequently occurring construction is NP-Complementation and the second is Subordinate Conjunction. And one could very well merge the third one, i.e., Relativization, with the first one. The total of these three rules is more than 50% out of simple accumulation of all rule occurrences. It means theoretically that a first filter consisting of these rules in the machine translation could process more than 50% of its task, and when applied to teaching Korean, students can learn the most heavily used patterns by mastering only three rules.

Next, Rules 1 to 10 occupy more than 90%, and those up to 15 already cover almost 99% of all occurrences. Of course, one must also consider the less frequent rules beyond the fifteenth in order to reflect the full complexity of the natural language.

Taking a more detailed look at this list, both number 2 and 5 are conjunctions ranked high in frequency. As for Pronominalization, personal

pronouns rank seventh, and indefinite pronouns fourteenth. Although many syntacticians have been attracted by Reflexivization, it is seventeenth, third from the bottom.

As for Negation, *an*-Negation is tenth, *mot*-Negation fifteenth, and *mal*-Negation is nineteenth. The total of these three negations is only 2.6%.

Passivization in long and short forms are thirteenth and fifteenth, respectively, and their total is 1.93%. However, including *-toy-* Passivization like *saenggak-toyda* 'be thought' which is ninth, whole passivization takes 4.63%. This percentage is comparatively low since the occurrence of passivization in English is about 10%. We may also compare other phenomena with various languages in order to investigate the idiosyncrasy of Korean.

Lastly, Causativization ranks very low, i.e., short form in the eighteenth, and long form in the twentieth. Despite researchers' fascination with this construction, Causativization occurs merely in 0.43%.

Admittedly, this is not an exhaustive list of syntactic phenomena in Korean. For instance, Equi-NP (or complement subject) deletion can be an added item to the above. In Korean syntax, Equi-NP deletion is usually defined as occurring in cases of **embedding**. However, when I looked for Equi-NP deletion according to this definition, I did not find any in the text. Thus I redefined it with the occurrence of identical NP exclusively in **conjunction**, and then I found the result that this 'modified' Equi-NP deletion ranks eighth in frequency. The reason why Equi-NP deletion did not show up in the initial search is perhaps due to the trend that embedding itself is rather rare in the written language, which I took as the main part of the corpus.

I think double subject, double objects and double negation should be checked. In addition, some phenomena mentioned in this list are not purely syntactic but morphological as well as stylistic. Therefore, one may process them in a separate stage.

In spite of incompleteness, these statistics already show a general trend, and I am sure that the overall picture will not change even if one or two additional rules are inserted.

5. Analysis of Various Kinds of Data

What has intrigued me more in the process of this investigation is that, depending on the genre of the text in question, the statistical results differ significantly. For instance, the legal language encompassing the Korean Constitution, university codes and some other codified prescriptions does not utilize pronominalization at all. I assume that the reason for this interesting phenomenon is to avoid any ambiguity in the legal language.

However, novels and poems make more use of pronominalization than the other genres of text. In addition, modified Equi-NP deletion in the **conjunction** sentence occurs more in this literary and artistic language. Novels contain more interrogative sentences than the other classes. Pluralization is also more frequent in novels. It reveals the simple habit of pluralization all through the constituents of a sentence in the colloquial style as observed very easily in casual speech.

Returning to the legal language, topicalization and *-toy-* passivization are very predominant over other elements. These facts are entitled to serve as all concrete bases for further studies in stylistics.

Before concluding, I would like to present the comparative description of frequency orders between phonetic/phonological rules and syntactic phenomena. Because these two areas of the grammar constitute the core part and reveal the tendency of their frequency. In fact, we will see interesting results from a comparison between syntactic and phonological cases.

The Functional Load of Phonetic/Phonological Rules:
A Quantitative Survey in Modern Korean

T: Phonetic Rules M: Phonemic Rules O: Others

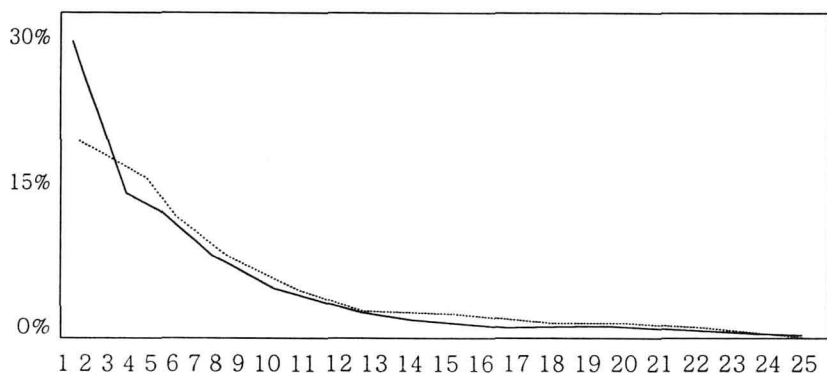
1. T Unrelease of voiced consonants		%
at syllable final positions	31.84	
2. T Voicing of voiceless obstruents		
in intervocalic positions	19.37	
		-51.21
3. M Vowel lengthening	10.00	
4. T Lateralization	9.78	
5. M Glottalization	6.43	

6. M Neutralization of syllable final obstruents	5.16	
7. T s-palatalization	3.71	
8. O syllable adjustment	3.10	
9. M Monophthongization	1.60	
10. T n-palatalization	1.42	
		-92.41
11. O Ø → we	0.95	
12. M Nasalization	0.82	
13. M l → n	0.70	
14. M Initial l & n(y) deletion	0.68	
15. M y-deletion after palatal consonants	0.67	
		-96.23
16. O ü → wi	0.66	
17. T l-palatalization	0.63	
18. M Aspiration	0.57	
19. M h-deletion	0.51	
20. M Liquidization	0.35	
		-98.95
21. O s → t	0.29	
22. O iy → i or i	0.29	
23. M n-insertion	0.26	
24. O iy → e	0.08	
25. M Consonant Cluster Simplification	0.07	
26. M t-palatalization	0.05	
27. M Decoronalization	0.03	
28. O ye → e	0.013	
29. M y-glidization	0.006	

Interestingly enough, there is similarity between syntactic and phonological cases. Only two or three most frequent rules on the top occupy more than 50%. The top 10 cover more than 90%, and the top 15 or 20 cover more than 99%.

Phonetic/Phonological Rules _____

Syntactic Phenomena



〈Figure 2〉

This parallelism in the distribution of frequency might be regarded as accidental, but we did not know this parallel distribution before quantification. This study reveals this hidden aspect of language: i. e., the concentration of functional load in a few frequent rules in both phonetics/phonology and syntax. It is quite similar to the prevailing use of VHS system among 'competing technologies' such as the Sony Betamax system (Brian Arthur 1994). The more frequent, the more used or vice versa.

The shape of these curves reminds us of Zipf's law or an indifference curve in economics. Even in many other unrelated disciplines, similar curve patterns can be observed in the investigation of frequency.

6. Conclusion

It is clear that the relative weight or importance of all different rules must be considered in the mechanical application of linguistic ideas. I have previously investigated the frequency of phonetic and phonological rules in order to support the economically optimal design of a speech recognition system. We can do the same thing here with syntactic rules by putting the highly ranked rules in the mainstream of processing while leaving minor rules in the periphery.

I hope to extend the scope of the present study to the area of "constraints" as focused in the framework of the Optimality Theory. That is, I would like to measure the various frequency levels of all different constraints introduced in studies of a certain language and to assign them a hierarchy of necessity and/or justification of introducing those constraints. In this way, we may reasonably limit the number of constraints that multiply to the point of negating the logical meaning of an identifiable constraint. Those constraints in the upper rank of frequency are of course necessary ones in the analysis of many languages. Those in the lower rank need more justification to be introduced in each study. This simple logic has led me to a strong conviction to study language from this new perspective.

It is known that this statistical type of approach is a new trend in the methodology of computational engineering. This has encouraged me to study hierarchical nature of linguistic structures by quantification through projects such as the above. Besides the possible application of the results of this study to linguistic engineering, other areas may benefit from this sort of study, e.g. checking the proper distribution of both kind of rules in textbooks of school grammar, conversation, and writing for first- and second-language acquisition.

References

- Arthur, W. Brian (1994) 'Competing Technologies, Increasing Returns, and Lock-In by Historical Small Events,' in W. Brian Arthur (ed.) *Increasing Returns and Path Dependence in the Economy*, The University of Michigan Press.
- King, Robert D. (1965) *Functional Load: its Measure and its Role in Sound Change*, Doctoral dissertation, University of Wisconsin, Madison.
- _____ (1967) 'Functional Load and Sound Change,' *Language* 43.4, 831~852.
- Lee, Sang-Oak (1977) 'Conspiracy in Korean Phonology Revisited: As Applied to Historical Data,' *Studies in the Linguistic Sciences* 7.2, 1~23.
- _____ (1989) 'A Glottometrical Study of Korean Lexicon,' *Harvard Studies in Korean Linguistics* III, Hanshin Publishing Co., Seoul, 159~166.

-
- (1990) 'On the Functional Load of Phonetic/Phonological Rules: A Quantitative Survey in Modern Korean,' *Language Research* 26.3, 441~467.
-
- (1997) 'A Survey on the Distribution of Phonetic/Phonological and Syntactic Phenomena Included in the Textbooks of Modern Korean Grammar,' *Kwanak Research of Korean Language and Literature* 22, 165~183.
- Meyerstein, R. S. (1970) *Functional Load: Descriptive Limitations, Alternatives of Assessment and Extensions of Application*, Janua Linguarum, Series minor 99, Mouton, The Hague.

Dept. of Korean Language and Literature
Seoul National University
San 56-1 Sinlim-dong, Kwanak-ku
Seoul 151-742, Korea